**Predicting survival rate in patients following heart failure using machine learning**

15/05/2021

Ho Yin Lam

RMIT University

s3889140@student.rmit.edu.au

Jingxuan Feng

RMIT University

s3843790@student.rmit.edu.au

We certify that this is all our own original work. If we took any parts from elsewhere, then

they were non-essential parts of the assignment, and they are clearly attributed in our

submission. We will show we agree to this honour code by typing "Yes": Yes.

**Table of Contents**

[**Abstract** 2](#_Toc72058025)

[**Introduction** 2](#_Toc72058026)

[**Methodology** 2](#_Toc72058027)

[**Results** 3](#_Toc72058028)

[**Discussion** 3](#_Toc72058029)

[**Conclusion** 3](#_Toc72058030)

[**References** 3](#_Toc72058031)

### **Abstract**

With the increased interest and usage of artificial intelligence by tech giants, they offer a varying range of services that improve our quality of life. These can come in the form of digital assistants such Alexa, Google home or embedded in self driving cars. We will also explore the feasibility of utilising machine learning to predict the likeliness patients surviving a heart failure based off a range of medical conditions and readings. We will investigate each attribute, what they represent and whether they play a part in helping to predict survival rate after heart failure.

### **Introduction**

We will be conduction our testing on the “Heart failure clinical records Data Set”. This dataset contains information on 299 patients from 2015. The patients included 105 females and 194 males, all ranging between 40 and 95. We will be investigating how the 13 features relating to the patients’ medical readings, conditions and lifestyle will affect the survival rate.

**Dataset Information** [4]

* Anaemia, high blood pressure, diabetes, gender, smoking and death event are all Boolean values. They will represent whether the feature is present or not.
* Age of the patients is measured in years.
* creatinine phosphokinase (CPK) measures the level of CPK enzyme in the blood. Measured in mcg/L.
* ejection fraction measures the percentage of blood exiting the heart after each contraction.
* Platelets is the number of platelets present in the blood. Measured in kilo platelets/mL.
* serum creatinine is the level of serum creatinine in the blood. Measured in mg/dL.
* Time is the number of days in the follow up period.

### **Methodology**

**Data Cleaning**

First, we will ensure the dataset is clean before any analysis or modelling is performed. We checked each column for errors as well as outliers. We found that several columns did, in fact include outliers but decided not to remove nor change any of these values as they are likely side effects of a medical symptom.

**Data Exploration**

Afterward, we proceeded to explore the data. We decided to standardise the way we explore the data when looking at the feature by itself. For all Boolean data, we used a pie chart to compare the percentage ‘yes’ vs ‘no’ for the feature. This allows us to see how prevalent each case is compared to each other.

For continuous data, we used a frequency histogram. This allows us to visualise how the values are distributed amongst the patients, as well as how they are grouped together.

When exploring relationships between pairs of attributes, we used a scatter diagram. The scatter diagram allows us to check for correlation between two features. This helps us locate figure out which features to pay closer attention to when investigating the likeliness of survival.

**Data Modelling**

We choose to use two classification models: Decision Tree and K-Nearest Neighbours from the sklearn library. We will use 70% of the dataset to train the model while the remaining 30% will be used to evaluate the accuracy of our model. As for the K value, I ran it through a for loop to check for the best k value to use, as well as graphing out the corresponding accuracy level for every other k value.

When selecting our features, I have used a correlation matrix and selected the features manually based off the results from the chi-square.

### **Results**

Discuss here.

### **Discussion**

Discuss here.

### **Conclusion**

Discuss here.

### **References**

* [1] ‘Correlation Matrix in Python - Practical Implementation - AskPython’. https://www.askpython.com/python/examples/correlation-matrix-in-python (accessed May 14, 2021).
* [2] ‘Plotting Correlation Matrix using Python’, *GeeksforGeeks*, Nov. 23, 2020. <https://www.geeksforgeeks.org/plotting-correlation-matrix-using-python/> (accessed May 14, 2021).
* [3] M. Ebrahim, ‘Python correlation matrix tutorial’, *Like Geeks*, Jun. 17, 2020. <https://likegeeks.com/python-correlation-matrix/> (accessed May 14, 2021).
* [4] Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, Muhammad Ali Raza, and Tanvir Ahmad, ‘Heart failure clinical records Data Set’, *UCI Machine Learning Repository*. <https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records> (accessed May 03, 2021).
* [5] D. Chicco and G. Jurman, ‘Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone’, *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 16, Feb. 2020, doi: [10.1186/s12911-020-1023-5](https://doi.org/10.1186/s12911-020-1023-5).
* [6] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, ‘Survival analysis of heart failure patients: A case study’, *PLOS ONE*, vol. 12, no. 7, p. e0181001, Jul. 2017, doi: [10.1371/journal.pone.0181001](https://doi.org/10.1371/journal.pone.0181001).